ASYMPTOTIC BEHAVIOR OF SOME FACTORIZATIONS OF RANDOM WORDS

PHILIPPE CHASSAING AND ELAHE ZOHOORIAN AZAD

Abstract. This paper considers the normalized lengths of the factors of the Lyndon decomposition of finite random words with n independent letters drawn from a finite or infinite totally ordered alphabet according to a general probability distribution. We prove, firstly, that the limit law of the lengths of the smallest Lyndon factors is a variant of the stickbreaking process. Convergence of the distribution of the lengths of the longest factors to a Poisson-Dirichlet distribution follows. Secondly, we prove that the distribution of the normalized length of the standard right factor of a random n-letters long Lyndon word, derived from such an alphabet, converges, when n is large, to:

$$\mu(dx) = p_1\delta_1(dx) + (1-p_1)\mathbf{1}_{[0,1)}(x)dx,$$

in which p_1 denotes the probability of the smallest letter of the alphabet.

1. Introduction

First, recall some general definitions from [Lot83, Reu93]. Let $\mathcal{A} = \{a_1, a_2, \dots\}$ be an ordered alphabet ($a_1 < a_2 < \dots$), finite or infinite, and let $\mathcal{A}^n$ be the corresponding set of n-letters long words. If $w \in \mathcal{A}^n$, write $w = w_1 \dots w_n$ and define $\tau w = w_2 \dots w_n w_1$. Then $\langle\tau\rangle = \{Id, \tau, \dots, \tau^{n-1}\}$ is the group of cyclic permutations of the letters of a word with length n. The orbit $\langle w\rangle$ of a word w under $\langle\tau\rangle$ is called a *necklace*. A word $w \in \mathcal{A}^n$ is called *primitive* if its necklace $\langle w\rangle$ has exactly n elements. In this case the necklace is said to be *aperiodic*. A *Lyndon word* is a primitive word that is minimal in its necklace, with respect to the lexicographic order. A word v is a *factor* of a word w if there exists two other words s and t, possibly empty, such that $w = svt$. If s (resp. t) is empty, v is a prefix (resp. a suffix) of w. A word v is a factor of a necklace $\langle w\rangle$ if v is a factor of some word $w' \in \langle w\rangle$.

The *standard right factor* v of a word w is its smallest proper suffix in the lexicographic order. The related factorization uv of w is often called the *standard factorization* of w. Both the standard right factor v and the corresponding prefix u (such that $w = uv$) of a Lyndon word are also Lyndon words. The standard factorization of a Lyndon word is the first step in the construction of some basis of the free Lie algebra over $\mathcal{A}$, due to Lyndon [Lyn54] (see for instance [Lot83] or [Reu93]).

On the other hand, according to [Lot83, Theorem 5.1.5],

2000 *Mathematics Subject Classification.* 60F05, 60C05, 68R15, 05A05.

Key words and phrases. Random word, Lyndon word, standard right factor, Lyndon factor, longest run, convergence in distribution, Poisson-Dirichlet distribution.

Theorem 1.1 (Lyndon)**.** *Any word $w \in \mathcal{A}^+$ has a unique factorization as a non-increasing product of Lyndon words:*

$$w = w_l w_{l-1} \dots w_2 w_1, \qquad w_i \in \mathcal{L}, \qquad w_l \geq w_{l-1} \geq \cdots \geq w_2 \geq w_1,$$

in which $\mathcal{A}^+$ is the set of nonempty words on alphabet $\mathcal{A}$ and $\mathcal{L}$ the set of Lyndon words on this alphabet.

If w is a Lyndon word, $w_1 = w$, else w_1 is the standard right factor of w. In this paper, we shall study the asymptotic behavior of probability distributions related to these factorizations.

$\mathtt{w}$ = aabb.aaabbbab.aaabb.a.a.a.a

FIGURE 1. A word w with 7 factors and a sequence of lengths : $\rho^{(20)}(w) = \frac{1}{20}\,(1,1,1,1,5,7,4,0,0\dots)$.

1.1. **Random words and random Lyndon words.** From now on, we consider a general probability distribution $(p_i)_{i\geq 1}$ on a set $\mathcal{A} = \{a_1 \leq a_2 \leq \dots\}$ of letters, and we assume, without loss of generality, that $0 < p_1 < 1$, i.e. the probability that a word has at least two letters does not vanish for n large. On the corresponding set of words,

$$\mathcal{A}^\star = \cup_{n\geq 0}\mathcal{A}^n = \{\varnothing\} \ \cup \ \left(\bigcup_{n\geq 1} \mathcal{A}^n\right),$$

we define the weight $p(w)$ of a word $w = a_{\ell_1(w)} a_{\ell_2(w)} \dots a_{\ell_n(w)}$ as

$$p(w) = p_{\ell_1(w)} p_{\ell_2(w)} \cdots p_{\ell_n(w)}.$$

The weight $p(.)$ defines a probability measure $\mathbb{P}_n$ on the set $\mathcal{A}^n$, through

$$\mathbb{P}_n(\{w\}) = p(w).$$

$\mathcal{P}_n$ (resp. $\mathcal{N}_n$, $\mathcal{L}_n$) denotes the set of n-letters long primitive words (resp. its complement, resp. the set of n-letters long Lyndon words). Then we define a probability measure $\mathbb{L}_n$ on $\mathcal{L}_n$, as follows

$$\mathbb{L}_n(\{w\}) = \lambda_n p(w),$$

in which $\lambda_n = 1/\mathbb{P}_n(\mathcal{L}_n) = n/\mathbb{P}_n(\mathcal{P}_n)$. The probability measure $\mathbb{L}_n$ has a trivial extension to $\mathcal{A}^n$ (setting $\mathbb{L}_n\left(\mathcal{L}_n^c\right) = 0$).

1.2. **Main results.** The sequence $\rho^{(n)}(w) = (\rho_{i,n}(w))_{i\geq 1}$ of normalized lengths of the Lyndon factors of a word $w \in \mathcal{A}^n$, with Lyndon factorization $w = w_\ell w_{\ell-1} \dots w_1$, is defined as follows:

$$\rho_{i,n}(w) = \begin{cases} \frac{|w_i|}{n} & \text{if } \ 1 \leq i \leq \ell \\ 0 & \text{if } \ i > \ell. \end{cases}$$

Thus $\rho_{i,n}(w)$ denotes the normalized length of the i-th smallest[1] Lyndon factor of w. Our first result describes the limit distribution, as n grows, of the sequence $\rho^{(n)}(w) = (\rho_{i,n}(w))_{i\geq 1}$, seen as a random variable on $(\mathcal{A}^n, \mathbb{P}_n)$. We have:

Theorem 1.2. *For a totally ordered alphabet with probability distribution p on its letters, $\rho^{(n)}$ converges in law, when $n \to \infty$, to the random sequence $\rho = (\rho_i)_{i\geq 1}$ whose law is defined by the law of ρ_1:*

$$\mu(dx) = p_1\delta_0(dx) + (1-p_1)\mathbf{1}_{(0,1]}(x)dx,$$

and the conditional law of ρ_i given $(\rho_1, \rho_2, \ldots, \rho_{i-1})$:

$$\mu^y(dx) = \begin{cases} p_1\delta_0(dx) + (1-p_1)\mathbf{1}_{(0,1]}(x)dx \ ; & y = 0 \\ \\ \frac{1}{1-y}\mathbf{1}_{(0,1-y]}(x)dx \ ; & y > 0, \end{cases}$$

in which y denotes $\rho_1 + \rho_2 + \cdots + \rho_{i-1}$.

In other words, if we set $s_i = 1-(\rho_1+\rho_2+\cdots+\rho_i)$, then $s = (s_i)_{i\geq 1}$ is a Markov chain starting from 1 at time 0, with transition probability

$$p(y, dx) = \begin{cases} p_1\delta_1(dx) + (1-p_1)\mathbf{1}_{(0,1]}(x)dx \ ; & y = 1 \\ \\ \frac{1}{y}\mathbf{1}_{(0,y]}(x)dx \ ; & y < 1. \end{cases}$$

The process s is a variant of the *stickbreaking process* [McC65, PPY92] related to the Poisson-Dirichlet(0,1) distribution, in which the first attempts to break the stick would fail (with probability p_1) and would produce a geometric number of fragments with size 0 at the beginning of the process, while for the stickbreaking process the transition probability $\tilde{p}(y, dx)$ is $\frac{1}{y}\mathbf{1}_{(0,y]}(x)dx$ for any $y \in [0,1]$. Of course, whence $\rho^{(n)}$ and ρ are rearranged in decreasing order, the small initial fragments are rejected at the end or, in the case of ρ, they disappear. Thus

Corollary 1.3. *The decreasing rearrangement of $\rho^{(n)}$ converges in law to the Poisson-Dirichlet(1) distribution.*

As regards the second result, for any Lyndon word $w \in \mathcal{L}_n$, let $R_n(w)$ denotes the length of its standard right factor, and set $r_n = R_n/n$. We have:

Theorem 1.4. *For a totally ordered alphabet with probability distribution p on its letters, the normalized lengths r_n of the standard right factor of a random n-letters long Lyndon word, when $n \to \infty$, converges in law to*

$$\mu(dx) = p_1\delta_1(dx) + (1-p_1)\mathbf{1}_{[0,1)}(x)dx,$$

where δ_1 denotes the Dirac mass on the point 1 and dx the Lebesgue measure on $\mathbb{R}$. As a consequence the moments of r_n converge to the corresponding moments of μ.

For instance, if p is the uniform distribution on q letters, then the limit law of the normalized length of the standard right factor of a random Lyndon word, is

$$\mu(dx) = \frac{1}{q}\delta_1(dx) + \frac{q-1}{q}\mathbf{1}_{[0,1)}(x)dx.$$

[1]In this paper, when applied to words, or factors of words and necklaces (and, sometimes, to factors with special properties, that we will call blocks), the adjectives "small" and "large" refer to the *lexicographic order* on words, while "short" and "long" refer to the *size* (number of letters) of factors.

1.3. **Context.** The Poisson-Dirichlet family of distribution was introduced by Kingman [Kin75]. This distribution arises as a limit for the size of components of decomposable structures in a variety of settings, as shown by Hansen [Han94] or Arratia et al. [ABT99].

When the distribution p is uniform on q letters, i.e.

$$p_k = \frac{1}{q}\, \mathbb{1}_{1\leq k\leq q},$$

the combinatorics of the Lyndon decomposition have connections with that of q-shuffles [BD92] and of monic polynomials of degree n over the finite field $GF(q)$, as explained in [GR93, DMP95]. When p is uniform, Corollary 1.3 is well known (cf. [ABT93, Han94]). Actually, for a uniform p, a precise description of the size of Lyndon factors in term of the standard Brownian motion is given in [Han93, ABT93]. Our contribution is twofold :

- in Theorem 1.2, we give a description of the sizes of factors *depending on their rank* in the decomposition. Obviously, the order of factors matters in the Lyndon decomposition of words, while it does not for polynomials or for shuffles ;
- the distribution p on letters is perfectly general (we only require more than one letter). As a consequence, to our knowledge, combinatorics do not provide closed form expressions for the distribution of sizes of factors. Thus, for a general p, we were not able to prove, or to disprove, the *conditioning relation* (cf. [ABT03, p. 2]) which is usually required for convergence to the Poisson-Dirichlet distribution in such settings.

Theorem 1.4 deals with random words conditioned to be Lyndon words. This Theorem is a first step, as were the papers [BCN05, MZA07], towards the study of the Lyndon tree, that describes the complexity of some algorithms computing bases of the free Lie algebra on $\mathcal{A}$. The line of the proof of Theorem 1.4 is the same as in [MZA07], where the case $q = 2$ was obtained. The proofs of some Lemmas and Theorems in this paper are similar to their analogs in [MZA07] : in this paper, we only give the proofs that are significantly different from the case $q = 2$. We refer the interested reader to [MZA07] for more remarks and explanations.

1.4. **Sketch of proofs.** Consider a nonrandom partition of $[0, 1]$ into a (large) number of subintervals with small widths, k of these subintervals being marked. After a random uniform shuffle of these subintervals, the positions $X = (X_i)_{1\leq i\leq k}$ of the marked subintervals is close to a k-sample of the uniform distribution on $[0, 1]$. More specifically, their Wasserstein distance is bounded by the maximal width of the subintervals, see Lemma 4.8.

To use this principle, one has to build a factorisation (partition) of the random word such that :

(1) the distribution of the random word is invariant under a random uniform shuffle of the factors (subintervals) ;
(2) the length of the factors is $o(n)$ (while the Lyndon factors are $\Theta(n)$) ;
(3) the marked factors are very small with respect to the lexicographic order, for they begin with large runs of the letter a_1.

Thus, the marked factors, called "long blocks", are strongly related to the Lyndon decomposition: they are prefixes of the longest Lyndon factors, and their positions, approximately uniform according to Lemma 4.8, govern the lengths of the Lyndon factors.

Section 2 is devoted to preliminary results on some statistics on runs. Specially useful is the observation that the length of the longest run of "a_1" is typically of order $\log_{1/p_1} n$.

In Section 3, we describe the partition of a random word with length n into distinct "long blocks" with length of order $\log_{\frac{1}{\beta}} n$, long blocks that begin with the longest runs of "a_1". We have to make sure that, with a high probability, the Lyndon property is preserved by permutation of these blocks.

Once these preliminary tasks are performed, we use the shuffling principle, Lemma 4.8, to prove the main results, Theorem 1.4 in Section 4, and Theorem 1.2 in Section 5.

2. Number of runs and length of the longest run

For $w \in \mathcal{P}_n$, let $\pi(w)$ denote the unique Lyndon word in the necklace of w. For $\alpha \geq 1$, we set

$$\|p\|_\alpha = \left(\sum_i p_i^\alpha \right)^{1/\alpha}.$$

The next Lemma allows to translate bounds on $\mathbb{P}_n$ into bounds on $\mathbb{L}_n$:

Lemma 2.1. *For $A \subset \mathcal{A}^n$, we have:*

$$\mid \mathbb{L}_n(A) - \mathbb{P}_n(\pi^{-1}(A)) \mid = \mathcal{O}\left(\|p\|_2^n \right).$$

Note that $\|p\|_1 = 1$, and that, under the assumption $\{0 < p_1 < 1\}$, $\|p\|_\alpha$ is strictly decreasing in α. Among other well known inequalities, we shall make use of $\|p\|_2 \leq \sqrt{\max p_i}$. We set

$$\beta = \max\{p_1, 1 - p_1\}.$$

For instance, the choice $A = \mathcal{A}^n$ leads to

$$\mid 1 - \mathbb{P}_n(\mathcal{L}_n) \mid = \mathcal{O}\left(\|p\|_2^n \right) = \mathcal{O}\left(\beta^n \right).$$

Due to Lemma 2.1, the asymptotic properties of statistics, such as the number of runs and the length of the longest runs, that behave nicely under cyclic permutations, are the same on random words or on random Lyndon words, and the preliminary results needed under $\mathbb{L}_n$ and under $\mathbb{P}_n$, for Theorems 1.2 and 1.4, are equivalent.

Proof. This proof rephrases in probabilistic terms some results of [Reu93, Section 7.1], to which the reader is referred for definitions. Let us define two sequences of subsets of $\mathcal{A}^n$,

$$\begin{aligned} \mathcal{A}_{n,k} &= \left\{ w \in \mathcal{A}^n \mid \exists v \in \mathcal{A}^k \text{ such that } w = v^{n/k} \right\}, \\ \mathcal{P}_{n,k} &= \mathcal{A}_{n,k} \setminus \left(\bigcup_{1 \leq i < k} \mathcal{A}_{n,i} \right), \end{aligned}$$

with probabilities $\nu_k = \mathbb{P}_n(\mathcal{A}_{n,k})$ and $\xi_k = \mathbb{P}_n(\mathcal{P}_{n,k})$, respectively. Clearly

$$\mathcal{A}_{n,n} = \mathcal{A}^n, \qquad \mathcal{P}_{n,n} = \mathcal{P}_n.$$

Also, if $k|n$, $(\mathcal{P}_{n,i})_{i|k}$ is a partition of $\mathcal{A}_{n,k}$ (else, both $\mathcal{A}_{n,k}$ and $\mathcal{P}_{n,k}$ are empty). Thus

$$\nu_k = \sum_{d|k} \xi_d,$$

and, by the Möbius inversion formula,

$$\xi_k = \sum_{d|k} \mu(d)\nu_{k/d}, \tag{1}$$

in which $\mu(d)$ denotes the Möbius function. On the other hand, when $k|n$,

$$\begin{aligned}
\nu_k &= \sum_{w\in\mathcal{A}_{n,k}} p(w) \\
&= \sum_{v\in\mathcal{A}^k} p(v)^{n/k} \\
&= \sum_{\sum_i r_i = k} \binom{k}{r_1, r_2, \dots} \left(p_1^{r_1} p_2^{r_2} \dots\right)^{n/k} \\
&= \|p\|_{n/k}^n .
\end{aligned}$$

Specializing (1) to $k = n$, we obtain

$$\mathbb{P}_n(\mathcal{P}_n) = \sum_{d|n} \mu(d)\, \|p\|_d^n . \tag{2}$$

Let the set of divisors of n be $\{1 < d_1 < d_2 < \cdots < d_\ell = n\}$. Then, by (2),

$$\begin{aligned}
\left|\mathbb{P}_n(\mathcal{P}_n) - 1 + \|p\|_{d_1}^n\right| &\le (\ell - 1)\, \|p\|_{d_2}^n \\
&\le (n-2)\, \|p\|_{d_2}^n ,
\end{aligned}$$

if n is not prime. Else $\mathbb{P}_n(\mathcal{P}_n) = 1 - \|p\|_{d_1}^n$. In any case, $\left|\mathbb{P}_n(\mathcal{P}_n) - 1 + \|p\|_{d_1}^n\right|$ is a $o\left(\|p\|_{d_1}^n\right)$, and, since $d_1 \ge 2$,

$$\mathbb{P}_n(\mathcal{P}_n^c) = \mathcal{O}\left(\|p\|_2^n\right). \tag{3}$$

Lemma 2.1 is a direct consequence of

$$\mathbb{L}_n(A) = \frac{\mathbb{P}_n(\pi^{-1}(A))}{\mathbb{P}_n(\mathcal{P}_n)},$$

and of (3). □

Definition 2.2. Set $\mathcal{B} = \{0, 1\}$. Now, let φ denote the morphism, from $\mathcal{A}^\star$ to $\mathcal{B}^\star$, that sends the letter a_1 on the digit 0, any other letter of $\mathcal{A}$ on the digit 1, and any word $w \in \mathcal{A}^n$ on a word $\varphi(w) \in \mathcal{B}^n$. We denote by $N_n(w)$ the number of runs in $\varphi(w)$, by $X_1(w), X_2(w), \dots, X_{N_n}(w)$ their lengths, by $N_n^{(e)}(w)$ (resp. $M_n^{(e)}(w)$) the number of runs of the digit $e \in \mathcal{B}$ in the word $\varphi(w)$ (resp. the maximal length of such runs).

Lemma 2.3 (Number of runs of the letter a_1)**.**

$$\mathbb{P}_n\left(N_n^{(0)} < \frac{p_1(1-p_1)}{2}\ n\right) \quad = \quad \mathcal{O}\left(n^{-1}\right),$$

and

$$\mathbb{L}_n\left(N_n^{(0)} < \frac{p_1(1-p_1)}{2}\ n\right) \quad = \quad \mathcal{O}\left(n^{-1}\right).$$

Proof. In the case $p_1 = 0.5$, $N_n^{(0)} - 1$ has a binomial distribution under $\mathbb{P}_n$ (see [MZA07, Lemma 4.1]), but this property is lost as soon as $p_1 \neq 0.5$. In this general case, we shall construct a random word $t_n(\omega)$ in $\mathcal{B}^n$ by truncation of an infinite word ω on the alphabet $\mathcal{B}$:

$$\omega = \omega_1\omega_2\omega_3\ldots \quad \longrightarrow \quad t_n(\omega) = \omega_1\omega_2\omega_3\ldots\omega_n.$$

In other words, ω is a Bernoulli process with parameter $1-p_1$. Let $\mathbb{P}$ denote the distribution of ω, an infinite product of Bernoulli distributions with parameter $1-p_1$.

We set $\xi(\omega) = 1 - \omega_1$. For an element ω of the (almost sure) subset Ω of infinite words that does not end with an infinite run, let $\eta(\omega) = (\eta_i(\omega))_{i\geq 1}$ (resp. $\theta(\omega) = (\theta_i(\omega))_{i\geq 1}$) denote the sequences of lengths of runs of the digit 0 (resp. 1) in ω. Under the probability measure $\mathbb{P}$,

- ξ, η and θ are independent,
- η is a sequence of independent geometric random variables with expectation $(1-p_1)^{-1}$,
- θ is a sequence of independent geometric random variables with expectation ${p_1}^{-1}$,
- ξ is a Bernoulli random variable with parameter p_1.

The proof of Lemma 2.3 relies on the fact that the distribution of the prefix t_n under the probability measure $\mathbb{P}$ is also the image of $\mathbb{P}_n$ under φ (in other terms, t_n and φ have the same distribution), and $N_n^{(\varepsilon)}$ will denote indifferently a statistic on $\varphi(w)$ or on $t_n(\omega)$.

For $k \geq 0$, set $S_k^\eta = \sum_{i=1}^k \eta_i$ and $S_k^\theta = \sum_{i=1}^k \theta_i$. Then $\left\{N_n^{(0)} \circ t_n(\omega) \leq k\right\}$ holds if and only if

$$\left\{\left\{\xi = 1 \text{ and } S_k^\eta(\omega) + S_{k-1}^\theta(\omega) \geq n\right\} \vee \left\{\xi = 0 \text{ and } S_k^\eta(\omega) + S_k^\theta(\omega) \geq n\right\}\right\}.$$

Thus, by Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}_n\left(N_n^{(0)} \leq k\right) &\leq \mathbb{P}\left(S_k^\eta + S_k^\theta \geq n\right)\\ &\leq \frac{Var(S_k^\eta + S_k^\theta)}{\left(n - \mathbb{E}\left[S_k^\eta + S_k^\theta\right]\right)^2}.\end{aligned}$$

With the choice $k = \frac{p_1(1-p_1)}{2}\ n$, we obtain

$$\mathbb{P}_n\left(N_n^{(0)} < \frac{p_1(1-p_1)}{2}\ n\right) = \mathcal{O}\left(n^{-1}\right).$$

In order to obtain this result for $\mathbb{L}_n$, note that for a primitive word w, we have $N_n^{(0)}(w) - 1 \leq N_n^{(0)}(\pi(w)) \leq N_n^{(0)}(w)$, then use Lemma 2.1 (see [MZA07, Lemma 4.1] for details). □

We also need some information about the length of the longest runs of "a_1" in a word $w \in \mathcal{A}^n$ and in its necklace $\langle w \rangle$, for, among these long runs, the longest is bound to be the prefix of the smallest Lyndon factor of w, or the prefix of the unique Lyndon word in $\langle w \rangle$. Also, the second longest is bound to be the prefix of the second smallest Lyndon factor of w or the prefix of the standard right factor of the Lyndon word in $\langle w \rangle$. Furthermore, if Theorem 1.4 is to be true, there should exist at least two long runs and, if Theorem 1.2 is to be true, the number of these long runs should grow indefinitely with n, like the number of Lyndon factors of the random word. These points are consequences of Theorems 1.4 and 1.2, but they are also some of the steps of the proofs of these Theorems. They are addressed by the next Lemmas. In this paper ε denotes a real number in $(0, 1/2)$.

Definition 2.4. *(Long runs and short runs)* We call *long run* (resp. *short run*) of $w \in \mathcal{A}^n$ a run of "a_1" with length at least (resp. smaller than) $(1-\varepsilon)\log_{1/p_1} n$. We denote by $H_n(w)$ the number of long runs of "a_1" in w.

Lemma 2.5 (Number of long runs)**.**

$$\mathbb{P}_n\left(H_n \geq \alpha n^{\varepsilon}\right) \quad = \quad 1 - \mathcal{O}\left(n^{-1}\right),$$

and

$$\mathbb{L}_n\left(H_n \geq \alpha n^{\varepsilon}\right) = 1 - \mathcal{O}\left(n^{-1}\right),$$

in which α is a constant smaller than $\frac{p_1(1-p_1)}{4}$.

Proof. We choose a positive constant $\alpha \in \left(0, \frac{p_1(1-p_1)}{4}\right)$, so that

$$c = -\alpha + \frac{p_1(1-p_1)}{4} > 0.$$

We assume that random words are produced the same way as in the proof of Lemma 2.3. We let, for $i \geq 1$,

$$B_i = \mathbf{1}_{\{\eta_i \geq (1-\varepsilon)\log_{1/p_1} n\}}.$$

Then, for $\omega \in \Omega$,

$$H_n\left(t_n(\omega)\right) \geq \sum_{1 \leq 2i-1 \leq N_n^{(0)}-1} B_{2i-1}(\omega) \qquad \text{if } \omega_1 = a, \tag{4}$$

and

$$H_n\left(t_n(\omega)\right) \geq \sum_{1 \leq 2i \leq N_n^{(0)}-1} B_{2i}(\omega) \qquad \text{if } \omega_1 = b. \tag{5}$$

Note also that, under $\mathbb{P}$, $(B_i)_{i\geq 1}$ is a Bernoulli process, and that its parameter $p(n,\varepsilon)$ satisfies $n^{\varepsilon-1} \leq p(n,\varepsilon) \leq n^{\varepsilon-1}/p_1$.

Thus relations (4) and (5), with Lemma 2.3, entail that, under $\mathbb{P}_n$, H_n is, roughly speaking, stochastically larger than the binomial distribution with parameters $\frac{p_1(1-p_1)}{4}\, n$ and $p(n,\varepsilon)$. More precisely, if $S_{p_1,n,\varepsilon}$ denotes a random variable distributed according to the binomial distribution with parameters $\pi_n = \left\lfloor \frac{p_1(1-p_1)n-2}{4} \right\rfloor$ and $p(n,\varepsilon)$, then

$$\mathbb{P}_n\left(H_n \leq \alpha n^{\varepsilon}\right) \quad \leq \quad \mathbb{P}_n\left(N_n^{(0)} < \frac{p_1(1-p_1)}{2}\, n\right) + \mathbb{P}\left(S_{p_1,n,\varepsilon} \leq \alpha n^{\varepsilon}\right).$$

But by the inequality of Okamoto [Oka58, Bol01], a binomial random variable $S_{n,p}$ with parameters n and p satisfies :

$$\mathbb{P}_n\left(|S_{n,p}-pn|\geq h\right)<\frac{(pqn)^{1/2}}{h}\exp\left(-h^2/2pqn\right).$$

As a consequence

$$\begin{aligned}\mathbb{P}\left(S_{p_1,n,\varepsilon}\leq \alpha n^{\varepsilon}\right) &\leq \mathbb{P}\left(S_{p_1,n,\varepsilon}-\pi_n p_{n,\varepsilon}\leq \left(\alpha n-\pi_n\right)n^{\varepsilon-1}\right)\\ &\leq \mathbb{P}\left(S_{p_1,n,\varepsilon}-\pi_n p_{n,\varepsilon}\leq -\left(4cn-6\right)n^{\varepsilon-1}/4\right)\\ &< c_n n^{-\varepsilon/2}\exp\left(-\frac{n^{\varepsilon}}{2{c_n}^2}\right),\end{aligned}$$

in which

$$\lim_n c_n=\sqrt{p_1(1-p_1)}/2c.$$

The first statement of the Lemma follows. For the proof of the second statement, we note that if w is a primitive word,

$$H_n\circ\pi(w)\geq H_n(w)-1, \tag{6}$$

with equality when w begins and ends with long runs. Together with Lemma 2.1, it entails that

$$\begin{aligned}\mathbb{L}_n\left(H_n\leq \alpha n^{\varepsilon}-1\right) &\leq \mathbb{P}_n\left(\{w\in\mathcal{P}_n, H_n\circ\pi(w)\leq \alpha n^{\varepsilon}-1\}\right)+\mathcal{O}\left(\beta^{n/2}\right)\\ &\leq \mathbb{P}_n\left(H_n\leq \alpha n^{\varepsilon}\right)+\mathcal{O}\left(\beta^{n/2}\right).\end{aligned}$$

and the Lemma follows. □

Recall that $M_n^{(1)}$ denote the length of the largest run of non-a_1 letters. We have:

Lemma 2.6 (Large values of the longest runs)**.** *Under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probabilities of the events $\left\{M_n^{(0)}\geq 2\log_{1/p_1}n\right\}$ and $\left\{M_n^{(1)}\geq 2\log_{1/(1-p_1)}n\right\}$ are $\mathcal{O}\left(n^{-1}\right)$.*

Proof. First, we give the proof for

$$A_n=\left\{M_n^{(1)}\geq 2\log_{1/(1-p_1)}n\right\}.$$

Again, we assume that the random words are produced the same way as in the proof of Lemma 2.3. For $y>0$, we have:

$$\begin{aligned}\mathbb{P}_n(M_n^{(1)}\leq y) &\geq \mathbb{P}(\forall i\in\{1,\ldots,n\},\ \theta_i\leq y)\\ &\geq \left(1-(1-p_1)^{\lfloor y\rfloor}\right)^n.\end{aligned}$$

Choosing $y=\left\lceil 2\log_{1/(1-p_1)}n\right\rceil-1$, we obtain that

$$\mathbb{P}_n\left(A_n\right)=\mathcal{O}\left(n^{-1}\right).$$

Note that for a primitive word w, we have

$$\begin{aligned}M_n^{(1)}\circ\pi(w) &= \max\{M_n^{(1)}(w),\left(X_1(w)+X_{N_n}(w)\right)\mathbb{1}_{w_1\neq a_1\neq w_n}\}\\ &\leq \max\left\{M_n^{(1)}(w),\left(X_1(w)\mathbb{1}_{w_1\neq a_1}+X_{N_n}(w)\mathbb{1}_{w_n\neq a_1}\right)\right\}\end{aligned}$$

Since $\mathbb{P}_n$ is invariant under words' reversal, (X_1, w_1) and (X_{N_n}, w_n) have the same probability distribution. Thus, from Lemma 2.1, we deduce that

$$\mathbb{L}_n(A_n) \leq 2\,\mathbb{P}_n\left(X_1 \mathbb{1}_{a_1 \neq w_1} \geq \log_{1/(1-p_1)} n\right) + \mathbb{P}_n(A_n) + \mathcal{O}\left(\|p\|_2^n\right).$$

which leads to the desired bound for $\mathbb{L}_n(A_n)$. Similar arguments hold for $M_n^{(0)}$. □

3. Long blocks of words and good words

We mentioned in Section 1.4 that the lengths of the Lyndon factors are governed by the positions of the longest runs of "a_1", but it was a rough simplification: as already explained in [MZA07], some of these longest runs have equal lengths so we need to compare longer factors beginning with these long runs, in order to decide which runs are the prefixes of the Lyndon factors. In this section, we prove that almost every word $w \in \mathcal{A}^n$ has a large number of long factors, that we call *long blocks*, sharing three properties:

Definition 3.1. The *long blocks* are the factors of w that:

- begin with a long run of "a_1",
- end just before another run of "a_1" (not necessarily the next run of "a_1"),
- have the smallest possible length larger than $1 + 3\log_{1/\beta} n$.

Our main argument is valid only on a subset of $\mathcal{A}^n$, the set $\mathcal{G}_n$ of *good words*:

Definition 3.2. A word $w \in \mathcal{A}^n$ is a *good word* if it satisfies the following conditions:

i. w has at least $\lfloor \alpha n^{\varepsilon} \rfloor$ long blocks,
ii. the long blocks of w do not overlap,
iii. if two long blocks have a common factor, its length is smaller than $3\log_{1/\beta} n$,
iv. for each long run of w there exists a long block beginning with this run,
v. $M_n^{(0)}(w) \leq 2\log_{1/p_1} n$,
vi. $M_n^{(1)}(w) \leq 2\log_{1/(1-p_1)} n$.

It turns out that $\mathcal{G}_n$ has a large probability:

Proposition 3.3. *Under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of $\mathcal{G}_n^c$ is $\mathcal{O}\left(n^{2\varepsilon-1}\log^2 n\right)$.*

For the proof of Proposition 3.3, we need a few lemmas:

Lemma 3.4. *Denote by E_n the set of words $w \in \mathcal{A}^n$ in which some $\lceil 3\log_{1/\beta} n \rceil$-letters long factor appears twice in the necklace $\langle w \rangle$, at two non-overlapping positions:*

$$E_n = \left\{ w \in \mathcal{A}^n \,|\, \exists (w', v, a, b) \in \langle w \rangle \times \mathcal{A}^{\lceil 3\log_{1/\beta} n \rceil} \times (\mathcal{A}^\star)^2 \text{ s.t. } w' = vavb \right\}.$$

Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of E_n is $\mathcal{O}\left(n^{-1}\right)$.

A key argument of the proof of the main results breaks down if some long block of the decomposition of a random word is a prefix of another long block, somewhere else in the word. In order to preclude that, we shall consider blocks with at least $\lceil 3\log_{1/\beta} n \rceil$ letters (at least thrice the length of the longest run(s)[2] of the letter a_1), and we shall use Lemma 3.4.

[2]The probability that there exists several runs with the same maximal length inside a n-letters long random word is non vanishing with n large, so $\log_{1/p_1} n$ characters would be too short.

Proof. We have

$$\mathbb{P}_n(E_n) = \mathcal{O}\left(n^2\ \beta^{3\log_{1/\beta} n}\right) = \mathcal{O}\left(n^{-1}\right), \tag{7}$$

in which n^2 is a bound for the number of positions of the pair of factors of w, and $\beta^{3\log_{1/\beta} n}$ is a bound for the conditional probability that the second factor is equal to the first factor, given the value of the first factor and the positions of the factors.

Due to Lemma 2.1, $\mathbb{L}_n(E_n)$ satisfies

$$\mid \mathbb{L}_n(E_n) - \mathbb{P}_n\left(\pi^{-1}(E_n)\right) \mid = \mathcal{O}\left(\beta^{n/2}\right),$$

and $\pi^{-1}(E_n) = E_n \cap \mathcal{P}_n \subset E_n$. □

Lemma 3.5 (Overlap of long blocks)**.** *Let F_n denote the set of words $w \in \mathcal{A}^n$ such that some factor of $\langle w \rangle$, $\lceil 7\log_{1/\beta} n \rceil$-letters long, contains two disjoint long runs. Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of F_n is $\mathcal{O}\left(n^{2\varepsilon-1}\ \log^2 n\right)$.*

Proof. The bound for $\mathbb{P}_n(F_n)$ has a factor n for the position of the $\lfloor 7\log_{1/\beta} n \rfloor$-letters long factor, a factor $49(\log_{1/\beta} n)^2$ for the positions of the 2 runs, and a factor $n^{2\varepsilon-2}$ for the probability of 2 disjoint runs at 2 specified positions. The proof extends to $\mathbb{L}_n$ by virtue of Lemma 2.1. □

Lemma 3.6. *Let I_n denote the set of words $w \in \mathcal{A}^n$ whose suffix of length $\lceil 6\log_{1/\beta} n \rceil$ contains a long run of "a_1". Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of I_n is $\mathcal{O}\left(n^{2\varepsilon-1}\ \log^2 n\right)$.*

Proof. We have

$$\mathbb{P}_n(I_n) \leq n^{\varepsilon-1}\ \lceil 6\log_{1/\beta} n \rceil.$$

The factor $\lceil 6\log_{1/\beta} n \rceil$, that will be explained in the next proof, counts the number of positions where such a long run could begin. The factor $n^{\varepsilon-1} = {p_1}^{(1-\varepsilon)\log_{1/p_1} n}$ is the probability that a long run begins at some given position. The result for $\mathbb{L}_n$ follows from Lemma 2.1, Lemma 3.5 and

$$\mathbb{P}_n\left(\pi^{-1}(I_n)\right) \leq \mathbb{P}_n(F_n).$$

□

Proof of Proposition 3.3. Consider the sets

$$V_n = \{ w \in \mathcal{A}^n \mid w \text{ satisfies v. and vi. and } H_n(w) \geq \alpha n^{\varepsilon} \}$$

and

$$\tilde{\mathcal{G}}_n = V_n \backslash \left(E_n \cup F_n \cup I_n\right).$$

Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of $\tilde{\mathcal{G}}_n^c$ is $\mathcal{O}\left(n^{2\varepsilon-1}\ \log^2 n\right)$, due to Lemmas 2.5, 2.6, 3.4 and 3.6. Let us prove that $\tilde{\mathcal{G}}_n \subset \mathcal{G}_n$.

Consider a word $w \in \tilde{\mathcal{G}}_n$, and in order to prove that w satisfies conditions **i.** and **iv.**, consider a k-letters long long run of w, $w = t\,a_1^k\,s$. Since $w \notin I_n$, $\lceil 3\log_{1/\beta} n \rceil$ characters after the beginning of this long run, we are still at least $\lceil 3\log_{1/\beta} n \rceil$ characters away from the end of the word w, and, since w satisfies condition **v.**, this long run is over at this point:

$$|t| + k \leq |t| + \lceil 3\log_{1/\beta} n \rceil \leq n - \lceil 3\log_{1/\beta} n \rceil.$$

On the other hand, we are away from the end of the corresponding long block by at most $\lceil (1-\varepsilon)\log_{1/p_1} n\rceil - 1 + M_n^{(1)}(w)$ characters: the length of a short run of the letter a_1 followed by a run of the letter(s) $\bar{a}_1$[3]. But, due to condition **vi.**,

$$\lceil (1-\varepsilon)\log_{1/p_1} n\rceil - 1 + M_n^{(1)}(w) \leq \lceil 3\log_{1/\beta} n\rceil,$$

so there is room enough for the long block to end before the end of the word. Thus to each long run is associated a long block, and w satisfies conditions **iv.**, and also **i.**, since $H_n(w) \geq \alpha n^{\varepsilon}$.

Let us check that $w \in \tilde{\mathcal{G}}_n$ satisfies the conditions **ii.** and **iii.**: a long block is shorter than $\lceil 6\log_{1/\beta} n\rceil$, due to conditions **v.** and **vi.**, so it can overlap with the next long block only if the 2 corresponding long runs are contained in some $\lceil 7\log_{1/\beta} n\rceil$-letters long factor, i.e. only it can overlap if $w \in F_n$. Finally if w satisfies **ii.** and fails to **iii.**, then $w \in E_n$. □

In the two following sections, we prove separately the main theorems, Theorem 1.4 and Theorem 1.2.

4. Proof of Theorem 1.4

First, let us draw some consequences of the definitions of the previous sections. Let $H_n(w)$ be the number of long blocks of a word $w \in \mathcal{A}^n$.

Proposition 4.1. *A good Lyndon word $w \in \mathcal{G}_n \cap \mathcal{L}_n$ satisfies the following points:*

1. *each long block, by definition a factor of $\langle w\rangle$, is also a factor of w,*
2. *long blocks are all distinct,*
3. *there exists a smallest (resp. a second smallest) long block,*
4. *given a sequence of long blocks, $(\zeta_i)_{1\leq i\leq k}$,* sorted in increasing lexicographic order*, and any sequence of words, $(v_i)_{1\leq i\leq k}$, the sequence $(\zeta_i v_i)_{1\leq i\leq k}$ is also sorted in increasing lexicographic order,*
5. *the smallest of the long blocks is a prefix of w,*
6. *either the second smallest of the long blocks is a prefix of the standard right factor of w, or $r_n(w) = 1 - \frac{1}{n}$.*

Proof. Item (1) follows from point **iv.** of Definition 3.2. Item (2) follows from point **iii.** of Definition 3.2. Item (3) follows from item (2) and from point **i.** of Definition 3.2, as soon as $\lfloor \alpha n^{\varepsilon}\rfloor \geq 2$, since $H_n(w) \geq \lfloor \alpha n^{\varepsilon}\rfloor$. For items (4) and (6), it can be useful to remember a basic fact about the lexicographic order: if two words t_1 and t_2 have prefixes, respectively s_1 and s_2, such that $s_1 < s_2$, it does not entail that $t_1 < t_2$. However, under the additional condition that s_1 is not a prefix of s_2, $s_1 < s_2$ entails $t_1 < t_2$. Thus item (4) fails only if some ζ_i is a prefix of some ζ_j, $i < j$. But this would violate point **iii.** of Definition 3.2. As a consequence of the definition of Lyndon words, w begins with one of the longest runs of a_1 in $\langle w\rangle$. This longest run is a prefix of some long block due to point **i.** of Definition 3.2. This, together with item (4), entails item (5).

For item (6), consider the two smallest long blocks, $\zeta_1 < \zeta_2$, in the necklace $\langle w\rangle$, and let k_1 and k_2 be the lengths of the runs they begin with: $\zeta_1 = a_1^{k_1} v_1$ and $\zeta_2 = a_1^{k_2} v_2$, in which the words v_i do not begin with the letter a_1. We know that

[3]from now on, we do not need to use Bernoulli processes anymore, so, rather than discussing runs of 0's and 1's in $\varphi(w)$, we shall get back to w and discuss, equivalently, runs of letters a_1 and $\bar{a}_1$, in which "*runs of* $\bar{a}_1$" stands for "*runs without any letter* a_1".

w begins necessarily with a long run. Thus w begins with a long block, necessarily ζ_1, for ζ_1 is not a prefix of any other long block (see the considerations leading to item (4)). The second smallest word in $\langle w\rangle$, $w_2 = \tau^r w$, begins with ζ_2 or with $a_1^{k_1-1}v_1$, but, since $a_1^{k_1-1}v_1$ or ζ_2 are at least $\lceil 3\log_{1/\beta} n\rceil$-letters long, they cannot be prefixes of each other, due to point **iii.** of Definition 3.2. Thus $r_n(w) = 1 - \frac{1}{n}$ if $a_1^{k_1-1}v_1 < \zeta_2$, and $r_n(w) = 1 - \frac{r}{n}$ if $a_1^{k_1-1}v_1 > \zeta_2$. □

By Definition 3.2 and Proposition 4.1, a good Lyndon word $w \in \mathcal{G}_n \cap \mathcal{L}_n$ has a unique decomposition

$$w = \beta_1 g_1 \beta_2 g_2 \dots \beta_{H_n(w)} g_{H_n(w)},$$

in which the β_i's are a permutation of the long blocks ζ_i's, now sorted with respect to their position inside w rather than in lexicographic order (but $\beta_1 = \zeta_1$). The g_i's fill the gaps between the β_i's, and, if not empty, they begin with the letter a_1, but do not end with letter a_1. As a consequence, if not empty, g_i has a unique decomposition

$$g_i = a_1^{j_1}\bar{a}_1^{k_1}a_1^{j_2}\bar{a}_1^{k_2}\dots a_1^{j_r}\bar{a}_1^{k_r},$$

where r and all the exponents are positive. This leads to the definition of *short blocks of good Lyndon words*:

Definition 4.2. The *short blocks*, denoted $(s_j)_j$, of a good Lyndon word $w \in \mathcal{G}_n \cap \mathcal{L}_n$ are the factors $a_1^{j_m}\bar{a}_1^{k_m}$ appearing in the unique decomposition of factors g_i. As a consequence, any $w \in \mathcal{G}_n \cap \mathcal{L}_n$ has a unique *block*-decomposition

$$w = Y_0(w)Y_1(w)\dots Y_{K_n(w)-1}(w)Y_{K_n(w)}(w),$$

in which the Y_i's stand either for a long block β_i or for a short block s_j.

Remark 4.3. Set $k_0 = \lceil (1-\varepsilon)\log_{1/p_1} n\rceil$. This decomposition of good Lyndon words can be seen as the decomposition of the elements of some submonoid of $\bigcup_{l\neq 1} a_1^{k_0}\mathcal{A}^\star a_l$, containing $\mathcal{G}_n \cap \mathcal{L}_n$, according to the *code*[4] κ_n defined below:

- κ_n contains any word $a_1^k a_{l_1}^{k_1} a_{l_2}^{k_2}\dots a_{l_r}^{k_r}$ such that $r \geq 1$, $1 \leq k < k_0$, $k_i \geq 1$, $l_i \neq 1$ for $1 \leq i \leq r$.
- κ_n contains the elements $t = uv$ of $\bigcup_{l\neq 1} a_1^{k_0}\mathcal{A}^\star a_l$, with $|u| = \lfloor 1 + 3\log_{1/\beta} n\rfloor$, such that t does not contain any factor $a_l a_1^{k_0}$, $l \neq 1$ (long blocks do not overlap), and such that, for $\ell \neq 1$, $a_l a_1$ is not a factor of v.

Remark 4.4. Note that the short blocks of some word $w \in \mathcal{G}_n \cap \mathcal{L}_n$ have less than

$$k_0 + 2\log_{1/(1-p_1)} n \leq 3\log_{(\frac{1}{\beta}\wedge\frac{1}{1-p_1})} n$$

letters, while the long blocks are not longer than

$$2 + (6-\varepsilon)\log_{(\frac{1}{\beta}\wedge\frac{1}{1-p_1})} n.$$

For a long block, count $\lfloor 2 + 3\log_{1/\beta} n\rfloor$ letters for the minimal size of a long block, plus eventually a run of "a_1" (a short one, due to point **ii.** of Definition 3.2, at most $\lceil -1 + (1-\varepsilon)\log_{1/p_1} n\rceil$ letters long starting before the $\lfloor 2 + 3\log_{1/\beta} n\rfloor$-limit) and a run of "$\bar{a}_1$", at most $\lceil -1 + 2\log_{1/(1-p_1)} n\rceil$ letters, due to point **v.** of Definition 3.2.

[4] we understand a code as defined in [Lot05, p. 7], for instance.

When the factors of the block-decomposition are sorted according to the lexicographic order, the long blocks β_i's turn out to be smaller than the s_i's, since they begin with longer runs of "a_1". By Proposition 4.1, the smaller of all these factors is $\beta_1 = Y_0(w)$. Let $J_n(w)$ denote the index of the second smaller factor, and let d_n denote its (normalized) position, defined by

$$d_n(w) = \frac{1}{n} \sum_{i=0}^{J_n(w)-1} |Y_i(w)|. \tag{8}$$

If $w \in a_1\mathcal{L}_{n-1}$ (this happens with probability $p_1 + o(1)$, according to (13)),

$$r_n(w) = 1 - 1/n,$$

while if $w \in \mathcal{G}_n \cap (\mathcal{L}_n \backslash a_1\mathcal{L}_{n-1})$, the second smaller block $Y_{J_n(w)}$, also a long block, is a prefix of the standard right factor, by Proposition 4.1, and

$$r_n(w) = 1 - d_n(w).$$

When $w \in \mathcal{G}_n$, both cases can be detected by inspection of the two smallest blocks.

Let $\mathbb{G}_n$ denote the conditional probability given $\mathcal{G}_n \cap \mathcal{L}_n$:

$$\mathbb{G}_n(A) = \frac{\mathbb{P}_n(A \cap \mathcal{G}_n \cap \mathcal{L}_n)}{\mathbb{P}_n(\mathcal{G}_n \cap \mathcal{L}_n)} = \frac{\mathbb{L}_n(A \cap \mathcal{G}_n \cap \mathcal{L}_n)}{\mathbb{L}_n(\mathcal{G}_n \cap \mathcal{L}_n)}.$$

Let $\mathbb{U}_A$ (resp. $\mathbb{U}_d$) denote the uniform probability distribution on a finite set A (resp. the uniform distribution on $[0,1]^d$, for a given integer d). As a first step in the proof of Theorem 1.4, we show that the distribution of d_n under $\mathbb{G}_n$ converges to $\mathbb{U}_1$ with respect to the *L_2-Wasserstein metric* $\mathcal{W}_2(.,.)$.

The L_2-Wasserstein metric $\mathcal{W}_2(.,.)$ is defined by

$$\mathcal{W}_2(\mu,\nu) \quad = \quad \inf_{\substack{\mathcal{L}(X)=\mu \\ \mathcal{L}(Y)=\nu}} \mathbb{E}\left[\|X-Y\|_2^2\right]^{1/2}, \tag{9}$$

in which μ and ν are probability distributions on $\mathbb{R}^d$, and $\|.\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$. Convergence of $\mathcal{L}(X_n)$ to $\mathcal{L}(X)$ with respect to $\mathcal{W}_2(.,.)$ entails convergence of X_n to X in distribution (see [Rac91]). The multidimensional case ($d \geq 1$) is needed for section 5.

As in [MZA07], the key point is the invariance of $\mathbb{G}_n$ under uniform random permutation of the blocks $\{Y_1(w), \ldots, Y_{K_n(w)}(w)\}$.

Notations 4.5. Let $\mathfrak{S}_n$ denotes the set of permutations of $\{1,\ldots,n\}$. For $w \in \mathcal{G}_n \cap \mathcal{L}_n$, and $\sigma \in \mathfrak{S}_{K_n(w)}$, we set

$$\sigma.w = Y_0(w)Y_{\sigma(1)}(w)\ldots Y_{\sigma(K_n(w))}(w),$$

and

$$C(w) = \{\sigma.w \ : \ \sigma \in \mathfrak{S}_{K_n(w)}\}.$$

Proposition 4.6. *Assume that $w \in \mathcal{G}_n \cap \mathcal{L}_n$, and $w' \in C(w)$: then $w' \in \mathcal{G}_n \cap \mathcal{L}_n$ and w' has the same multiset of blocks as w (it has the same blocks, with the same multiplicity). As a consequence, for $w, w' \in \mathcal{G}_n \cap \mathcal{L}_n$, either $C(w) = C(w')$ or $C(w) \cap C(w') = \emptyset$.*

This follows directly from Definition 3.2 and the definition of a code. Let $\mathcal{C}_n = \{C(w)\,;\ w \in \mathcal{G}_n \cap \mathcal{L}_n\}$, and let $\mathfrak{C}_n$ denote the σ-algebra generated by $\mathcal{C}_n$. Also, let $X(w) = (X_i(w))_{i\geq 0}$ be the sequence of blocks of w sorted in increasing lexicographic order, ended by an infinite sequence of empty words, and let $\Xi(w) = (\Xi_i(w))_{i\geq 0}$ be the corresponding sequence of lengths.

Corollary 4.7. *The weight $p(.)$, X, Ξ, H_n and K_n are $\mathfrak{C}_n$-measurable, and*

$$\mathbb{G}_n = \sum_{C\in\mathcal{C}_n} \frac{\mathrm{Card}(C)\ p(C)}{\mathbb{P}_n(\mathcal{G}_n \cap \mathcal{L}_n)}\ \mathbb{U}_C.$$

Given that $w \in C$, the ranks of the blocks $(X_i)_{1\leq i\leq K_n(C)}$ are uniformly distributed.

Proof. The weight $p(w)$ depends only on the number of letters a_1, a_2, ... that w contains, not on the order of the letters in w, so that $p(.)$ is constant on each $C \in \mathcal{C}_n$: thus, under $\mathbb{G}_n$, the conditional distribution of w given that $w \in C$ is $\mathbb{U}_C$. As a consequence of Proposition 4.6, $\mathcal{C}_n$ is a partition of $\mathcal{G}_n \cap \mathcal{L}_n$, so the relation in Corollary 4.7 is just the decomposition of $\mathbb{G}_n$ according to its conditional distributions given $\mathcal{C}_n$. □

Due to the previous considerations, the general result below can be applied, in this section, to prove the asymptotic uniformity of d_n.

Lemma 4.8. *Let $\mathcal{W}_2(.,.)$ denote the L_2-Wasserstein metric on $\mathbb{R}^k$. Consider a random partition of $[0,1)$ into $\ell+2$ intervals $([a_i(\omega), b_i(\omega)))_{0\leq i\leq \ell+1}$, with respective (non-random) widths $(x_i)_{0\leq i\leq 1+\ell}$ ($x_i \geq 0$, $\sum_i x_i = 1$), sorted according to a random permutation $\omega \in \mathfrak{S}_\ell$, meaning that, for $1 \leq j \leq \ell$:*

$$a_j(\omega) = x_0 + \sum_{\substack{i:\ 1\leq i\leq \ell,\\ \text{and } \omega(i)<\omega(j)}} x_i, \qquad b_i(\omega) = x_0 + \sum_{\substack{i:\ 1\leq i\leq \ell,\\ \text{and } \omega(i)\leq\omega(j)}} x_i.$$

and

$$[a_0(\omega), b_0(\omega)] = [0, x_0], \qquad [a_{\ell+1}(\omega), b_{\ell+1}(\omega)] = [1 - x_{\ell+1}, 1].$$

Set $\tilde{a}_k = (a_1, \ldots, a_k)$; $1 \leq k \leq \ell$. Then

$$\mathcal{W}_2(\tilde{a}_k, \mathbb{U}_k) \leq \sqrt{\tfrac{k}{3} \sum_{i=1}^{\ell} x_i^2}.$$

Proof. The proof is similar to the proof of [MZA07, Lemma 6.3], which is the special case $k = 1$ of Lemma 4.8. As in the proof of [MZA07, Lemma 6.3], we rather define the random permutation ω and the sequence (a_i) with the help of a sequence of i.i.d. uniform random variables $(U_1, \ldots, U_\ell)$:

$$a_j = x_0 + \sum_{\substack{i:\ 1\leq i\leq \ell,\\ \text{and } U_i<U_j}} x_i.$$

Among the many couplings between $\tilde{a}_k$ and $\mathbb{U}_k$, this special one provides the desired bound on the Wasserstein distance. Actually, conditioning given U_j, $1 \leq j \leq k$, we

obtain

$$\begin{aligned}\mathbb{E}\left[(U_j-a_j)^2\right] &= \mathbb{E}\left[\left(x_0(U_j-1)+\sum_{i=1}^{\ell}x_i\left(U_j-1_{\{U_i<U_j\}}\right)\ +x_{\ell+1}U_j\right)^2\right]\\ &= \mathbb{E}\left[(1-U_j)^2\right]\ x_0^2+\mathbb{E}\left[U_j^2\right]\ x_{\ell+1}^2+\mathbb{E}\left[U_j(1-U_j)\right]\ \sum_{i=1}^{\ell}x_i^2\\ &= \tfrac{1}{3}\ \left(x_0^2+x_{\ell+1}^2\right)\ +\ \tfrac{1}{6}\ \sum_{j=1}^{k}\ x_j^2.\end{aligned}$$

We struggle with the idea that such computations are new. Actually the argument can be adapted (taking the x_i's in $\{0,1/n\}$) to compute the L_2 distance $(\sum t(1-t))/n$ between an evaluation $F_n(t)$ of the empirical distribution function and t, cf. [SW09, Ch. 3.1, p.85, display (3)]. □

We shall need the full generality of Lemma 4.8 in Section 5. In this section, we specialize Lemma 4.8 to $k=1$. If ν_n denotes the distribution of d_n under $(\mathcal{G}_n\cap\mathcal{L}_n,\mathbb{G}_n)$, we deduce that:

Theorem 4.9 (Position of the second smallest block)**.**

$$\mathcal{W}_2\left(\nu_n,\mathbb{U}_1\right)=\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right).$$

As a consequence, under $\mathbb{G}_n$, the moments of d_n converge to the corresponding moments of $\mathbb{U}_1$.

Proof. The proof of [MZA07, Theorem 6.4] holds step by step: if ν_C is the conditional distribution of $d_n(w)$ given that $w\in C$, then ν_C is also the image of the uniform probability on $\mathfrak{S}_{K_n(w)}$ by the application $\sigma\longmapsto d_n(\sigma.w)$. Thus Lemma 4.8 and (8) lead to

$$\mathcal{W}_2\left(\nu_C,\mathbb{U}_1\right)\le\frac{1}{n}\ \sqrt{\sum_i\Xi_i^2}.$$

Then, Corollary 4.7 entails

$$\mathcal{W}_2\left(\nu_n,\mathbb{U}_1\right)\le\frac{1}{n}\ \sqrt{\mathbb{E}\left[\sum_i\Xi_i^2\right]},$$

and we conclude with the help of $\sum_i a_i^2\le\left(\sum_i a_i\right)\times\max_i a_i$, and of Remark 4.4. □

As in [MZA07, Theorem 6.5], asymptotic independence between $\mathfrak{C}_n$ and d_n holds under $\mathbb{G}_n$: for a $\mathfrak{C}_n$-measurable $\mathbb{R}$-valued statistic W_n with probability distribution χ_n,

$$\mathcal{W}_2\left((W_n,d_n),\chi_n\otimes\mathbb{U}_1\right)=\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right). \tag{10}$$

In order to prove Theorem 1.4, let μ_n (resp. $\tilde{\mu}_n$) denote the image of $\mathbb{L}_n$ (resp. of $\mathbb{G}_n$) by r_n. Set

$$\mathcal{L}_n^1=(\mathcal{G}_n\cap\mathcal{L}_n)\cap a_1\mathcal{L}_{n-1}=\mathcal{G}_n\cap a_1\mathcal{L}_{n-1},\quad\text{and}\quad\mathcal{L}_n^2=(\mathcal{G}_n\cap\mathcal{L}_n)\backslash\mathcal{L}_n^1.$$

We remark that:

- **i.** if $w \in \mathcal{L}_n^1$, $r_n(w) = 1 - \frac{1}{n}$ holds true[5] ;
- **ii.** if $w \in \mathcal{L}_n^2$, $r_n(w) = 1 - d_n(w)$;
- **iii.** when $w \in \mathcal{L}_n \backslash (\mathcal{G}_n \cap \mathcal{L}_n)$, the crude bound $0 \leq r_n(w) \leq 1$ will prove to be more than sufficient for our purposes.

First, the conditional law $\tilde{\nu}$, given A, of a bounded r.v. X, defined on a probabilistic space Ω, is Wasserstein-close to its unconditional law ν, if A is close to Ω. More precisely

$$\mathcal{W}_2\left(\nu, \tilde{\nu}\right) \leq 2\,\mathbb{P}\left(\Omega \backslash A\right)^{1/2}\, \left\|X\right\|_\infty . \tag{11}$$

As a consequence, point **iii.**, together with Proposition 3.3, entails that

$$\mathcal{W}_2(\mu_n, \tilde{\mu}_n) = \mathcal{O}\left(n^{-1/2+\varepsilon}\ \log n\right).$$

Thus we shall now work on $\mathcal{G}_n \cap \mathcal{L}_n$, under $\mathbb{G}_n$, for μ_n has the same asymptotic behaviour as $\tilde{\mu}_n$.

On $\mathcal{G}_n \cap \mathcal{L}_n$, we have, according to points **i.** and **ii.**,

$$r_n = f_n\left(d_n, \mathbf{1}_{\mathcal{L}_n^2}\right) = (1 - d_n)\mathbf{1}_{\mathcal{L}_n^2} + \left(1 - \frac{1}{n}\right)(1 - \mathbf{1}_{\mathcal{L}_n^2}).$$

The $\mathfrak{C}_n$-measurability of $\mathcal{L}_n^2$ (see [MZA07, Section 7] for more details) and relation (10) entails that

$$\mathcal{W}_2\left((\mathbf{1}_{\mathcal{L}_n^2}, d_n), \chi_n \otimes \mathbb{U}_1\right) \quad = \quad \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right). \tag{12}$$

in which χ_n denotes the probability distribution of $\mathbf{1}_{\mathcal{L}_n^2}$. Thus, there exists a probability space, and, defined on this probability space, a couple (W_n, U) with distribution $\chi_n \otimes \mathbb{U}_1$, and a copy[6] of $(\mathbf{1}_{\mathcal{L}_n^2}, d_n)$ whose $\mathbb{L}^2$ distance satisfies

$$\left\|\mathbf{1}_{\mathcal{L}_n^2} - W_n\right\|_2^2 + \left\|d_n - U\right\|_2^2 = \mathcal{O}\left(\frac{\log n}{n}\right).$$

Set

$$\tilde{r}_n = (1 - U)W_n + \left(1 - \frac{1}{n}\right)(1 - W_n).$$

The inequality

$$\left|f_n\left(d, w\right) - f_n\left(d', w'\right)\right|^2 \leq 2\left(\left|d - d'\right|^2 + \left|w - w'\right|^2\right),$$

that holds for $(w, w', d, d') \in [0, 1]^4$, entails that

$$\mathcal{W}_2\left(\tilde{\mu}_n, \tilde{r}_n\right) \quad = \quad \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right).$$

[5]Actually, $r_n(w) = 1 - \frac{1}{n}$ holds true if $w \in a_k\mathcal{L}_{n-1}(a_k, a_{k+1}, \ldots, a_n)$, but, since $w \in \mathcal{G}_n$, w contains at least one occurrence of the letter a_1, which precludes $w \in a_k\mathcal{L}_{n-1}(a_k, a_{k+1}, \ldots, a_n)$ for $k \geq 2$.

[6]denoted $(\mathbf{1}_{\mathcal{L}_n^2}, d_n)$ for sake of economy.

Finally, using an optimal coupling $\left(W_n, \widehat{W}_n\right)$ in which $\widehat{W}_n$ is a Bernoulli random variable with expectation $1-p_1$, independent of U, set

$$\widehat{r}_n \quad = \quad (1-U)\widehat{W}_n + \left(1-\frac{1}{n}\right)(1-\widehat{W}_n).$$

As above, we obtain easily

$$\begin{aligned}\mathcal{W}_2\left(\tilde{r}_n, \widehat{r}_n\right) &\leq \mathcal{W}_2\left(W_n, \widehat{W}_n\right)\\ &\leq \sqrt{|\mathbb{G}_n(\mathcal{L}_n^2) - (1-p_1)|}.\end{aligned}$$

Also

$$(1-U)\widehat{W}_n + (1-\widehat{W}_n) = \widehat{r}_n + \frac{1}{n}(1-\widehat{W}_n)$$

has distribution μ. Thus

$$\mathcal{W}_2\left(\widehat{r}_n, \mu\right) \leq \frac{1}{n}.$$

Now

$$\mathbb{P}_n(a_1\mathcal{L}_{n-1}) - \mathbb{P}_n(\mathcal{L}_n \backslash \mathcal{G}_n) \leq \mathbb{P}_n(\mathcal{L}_n^1) \leq \mathbb{P}_n(a_1\mathcal{L}_{n-1}).$$

So by Proposition 3.3 and the fact that $\mathbb{P}_n(\mathcal{L}_n) = \frac{1}{n}(1-O(\beta^{n/2}))$, we obtain

$$\left|\mathbb{G}_n(\mathcal{L}_n^1) - p_1\right| = \mathcal{O}\left((\log n)^2\ n^{2\varepsilon-1}\right) \tag{13}$$

and

$$\mathcal{W}_2(\tilde{r}_n, \hat{r}_n) = \mathcal{O}\left(n^{-1/2+\varepsilon}\log n\right).$$

With (12), this yields

$$\mathcal{W}_2(\mu_n, \mu) = \mathcal{O}\left(n^{-1/2+\varepsilon}\log n\right).$$

Since $0 \leq r_n \leq 1$, convergence of moments follows. □

5. Proof of Theorem 1.2

According to Definitions 3.1 and 3.2, for $w \in \mathcal{G}_n$, the number of long blocks and the number of long runs are the same. Thus w has a unique decomposition

$$w = \bar{g}_1\beta_1 g_2\beta_2 \dots \beta_{H_n(w)}\bar{g}_{H_n(w)},$$

in which the β_i's are long blocks and $(\bar{g}_i)_{i\in\{1,H_n(w)\}}$ and $(g_i)_{i\in\{2,\dots,H_n(w)-1\}}$ are some words in $\mathcal{A}^\star$. The factors $\bar{g}_1$ and $\bar{g}_{H_n(w)}$ have a unique factorization :

$$\bar{g}_1 = \bar{a}_1^k a_1{}^{j_1}\bar{a}_1^{k_1}\dots a_1{}^{j_h}\bar{a}_1^{k_h} := \bar{a}_1^k g_1$$

and

$$\bar{g}_{H_n(w)} = a_1^{j_1}\bar{a}_1^{k_1}a_1^{j_2}\bar{a}_1^{k_2}\dots a_1^{j_{h'}}\bar{a}_1^{k_{h'}}a_1^j := g_{H_n(w)}a_1^j,$$

in which "$\bar{a}_1^k$" denotes a run of k letters that does not contain the letter "a_1", and h, h' and all powers are positive or zero. Then if a factor $g_{i;\ i\in\{1,\dots,H_n(w)\}}$ is non empty, it has a unique decomposition

$$g_i = a_1^{j_1}\bar{a}_1^{k_1}a_1^{j_2}\bar{a}_1^{k_2}\dots a_1^{j_r}\bar{a}_1^{k_r},$$

in which r and all the exponents are positive. Now let us define the *short blocks* of good words :

Definition 5.1. The short blocks, denoted $(s_j)_j$, of a good word $w \in \mathcal{G}_n$ are the factors $a_1^{j_m}\bar{a}_1^{k_m}$ appearing in the unique decomposition of factors g_i.

As a consequence, any word $w \in \mathcal{G}_n$ has a unique block-decomposition

$$w = \bar{a}_1^{k(w)} Y_1(w) \dots Y_{K'_n(w)-1}(w) Y_{K'_n(w)}(w) a_1^{L_n(w)},$$

in which $k \geq 0$, $L_n(w) \geq 0$ and the Y_i's are either long blocks, or short blocks. Let $J_{i,n}(w)$, $1 \leq i \leq H_n(w)$, denote the index of the i-th smallest block of $w \in \mathcal{G}_n$: since $i \leq H_n(w)$, $Y_{J_{i,n}}$ has to be a long block. Let $d_{i,n}(w)$, $1 \leq i \leq H_n(w)$, denote the normalized position of $Y_{J_{i,n}}(w)$ in w, defined as the ratio $|u|/|w|$, in which w has the factorization $w = uY_{J_{i,n}}(w)v$. The normalized position $d_{i,n}(w)$ is given by the formula :

$$d_{i,n}(w) = \frac{1}{n} \left(k(w) + \sum_{j=1}^{J_{i,n}(w)-1} |Y_j(w)| \right) \; ; \quad i = 1, \dots, H_n.$$

For a word $\omega \in \mathcal{A}^n$, it is convenient to complete the sequence $(d_{k,n}(\omega))_{1 \leq k \leq H_n(\omega)}$ by an infinite sequence of 0's. For a word $\omega \in \mathcal{G}^n$, this is not much of a perturbation, since the original sequence is very long : according to Lemma 2.5, the probability that $H_n(\omega)$ is smaller than αn^{ε} vanishes.

Let $\widehat{\mathbb{G}}_n$ denote the conditional probability given $\mathcal{G}_n$:

$$\widehat{\mathbb{G}}_n(A) = \frac{\mathbb{P}_n(A \cap \mathcal{G}_n)}{\mathbb{P}_n(\mathcal{G}_n)}.$$

By arguments similar to those in Section 4, we obtain that for any $k \geq 1$ the sequence of random variables $(d_{i,n})_{1 \leq i \leq k}$ is, under $\widehat{\mathbb{G}}_n$, asymptotically uniform on $[0,1]^k$. Once again, the key point is the invariance of $\widehat{\mathbb{G}}_n$ under uniform random permutations of the blocks Y_i : let $\sigma \in \mathfrak{S}_{K'_n(w)}$ act on w by permutation of blocks :

$$\sigma.w = \bar{a}_1^k Y_{\sigma(1)}(w) \dots Y_{\sigma(K'_n(w))}(w) a_1^{L_n(w)}.$$

The action is slightly different from the action defined at Section 4, for the decomposition is different, and in addition to the prefix $\bar{a}_1^k$, an eventual suffix $a_1^{L_n(w)}$ is also left untouched by the permutation. Let

$$C(w) = \{\sigma.w \; : \; \sigma \in \mathfrak{S}_{K'_n(w)}\}$$

denote the orbit of w under that action, and let $\mathfrak{C}'_n$ the σ-algebra generated by $\mathcal{C}'_n = \{C(w)\,;\; w \in \mathcal{G}_n\}$. The proof of the next theorem is similar of the proof of Theorem 4.9 in Section 4.

Theorem 5.2 (Positions of the k first smallest blocks)**.** *Let $\tilde{\nu}_{k,n} = (\nu_{i,n})_{1 \leq i \leq k}$ be the distribution of $\tilde{d}_{k,n} = (d_{i,n})_{1 \leq i \leq k}$ under $(\mathcal{G}_n, \widehat{\mathbb{G}}_n)$. We have*

$$\mathcal{W}_2(\tilde{\nu}_{k,n}, \mathbb{U}_k) = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right).$$

So far we saw that the normalized positions of the H_n smallest blocks are asymptotically independent and uniformly distributed on $[0,1]$. These H_n blocks are the prefixes of the H_n smallest words in the necklace of w. However, some of these H_n blocks are not prefixes of Lyndon factors of w, but if the sequence $i \to J_{i,n}$ is decreasing. For instance, in a word containing 9 long blocks with lexicographic

ranks going from 1 to 9, the blocks could be placed along the word in the following way :

$$\ldots 4 \ldots 8 \ldots 3 \ldots 5 \ldots 7 \ldots 1 \ldots 9 \ldots 2 \ldots 6 \ldots .$$

In this example, the long blocks which are prefixes of Lyndon factors of this word are the blocks 1, 3 and 4, those whose ranks constitute records of the sequence $\underline{4}8\underline{3}5\underline{7}1926$: the long blocks with ranks 9, 2 and 6 are immersed in the first Lyndon factor starting with $Y_{J_{1,n}}$, and the long blocks with ranks 5 and 7 are immersed in the second Lyndon factor, that starts with $Y_{J_{3,n}}$. Note that the largest (and shortest) Lyndon factors, that do not begin with long blocks, do not appear in this list of H_n factors, but, as a consequence of Theorem 1.2, the total length of these largest factors is $o(n)$: once normalized by n, their length does not contribute to the asymptotic behavior of the factorization.

By invariance of $\widehat{\mathbb{G}}_n$ under uniform random permutations of the blocks Y_i, the sequence of ranks of the long blocks is, conditionally given that $H_n = k$, a random uniform permutation of $\mathfrak{S}_k$. Thus the conditional distribution of the number Λ_n of Lyndon factors obtained this way, given that $H_n = k$, has the same law as the number of records (or of cycles) of a uniform random permutation in $\mathfrak{S}_k$ (see [ABT03, Ch. 1] or [Lot02, Ch. 11]), with generating function

$$\frac{1}{k!}\, x(x+1)(x+2)\ldots(x+k-1) = \frac{1}{k!} \sum_{0\le j\le k} \begin{bmatrix} k \\ j \end{bmatrix} x^j,$$

in which $\begin{bmatrix} k \\ j \end{bmatrix}$ is a Stirling number of the first kind. We can thus describe the conditional law of Λ_n as follows : consider a sequence $B = (B_i)_{i\ge 1}$ of independent Bernoulli random variables with respective parameters $1/i$, B and H_n being independent. Set

$$S_n = \sum_{1\le i\le n} B_i$$

and

$$\tilde{\Lambda}_n = S_{H_n} = \sum_i B_i \, \mathbb{1}_{1\le i\le H_n}.$$

Then Λ_n and $\tilde{\Lambda}_n$ have the same distribution, and we shall use the notation Λ_n for both of them. The following lemma insures that with a probability close to 1, Λ_n is at least of order $\log n$.

Lemma 5.3.

$$\mathbb{P}_n(\Lambda_n \le \varepsilon \log n/3) = \mathcal{O}\left(\frac{1}{\log n}\right).$$

Proof. The m-th harmonic number has the asymptotic expansion

$$\sum_{i=1}^{m} 1/i \;=\; H_m \;=\; \ln m + \gamma + \frac{1}{2m} - \frac{1}{12m^2} + \ldots,$$

in which γ is the Euler-Mascheroni constant (see [CG96]). We have

$$\mathbb{E}(S_n) = H_n \qquad \text{and} \qquad \mathrm{Var}(S_n) = H_n - \sum_{i=1}^{n} \frac{1}{i^2}.$$

By Lemma 2.5 :

$$\mathbb{P}_n\left(\Lambda_n < \varepsilon \log n/3\right) \;\le\; \mathbb{P}_n\left(\Lambda_n < \varepsilon \log n/3 \mid H_n \ge \alpha n^{\varepsilon}\right) + \mathcal{O}\left(n^{-1}\right).$$

In addition

$$\begin{aligned}\mathbb{P}_n\left(\Lambda_n < \varepsilon \log n/3 \mid H_n \geq \alpha n^{\varepsilon}\right) &\leq \mathbb{P}_n\left(S_{\alpha n^{\varepsilon}} < \varepsilon \log n/3\right) \\ &\leq \mathbb{P}_n\left\{|S_{\alpha n^{\varepsilon}} - \mathbb{E}(S_{\alpha n^{\varepsilon}})| \geq \varepsilon \log n/2\right\} \\ &= \mathcal{O}\left(\frac{1}{\log n}\right),\end{aligned}$$

in which the second inequality holds true for n large enough, and the last equality follows from the Bienaymé-Chebyshev inequality. □

Let L denote a geometric random variable with parameter $1-p_1$, such that, for $k \geq 0$,

$$\mathbb{P}(L = k) = p_1^k(1-p_1),$$

and let $U = (U_k)_{k\geq 1}$ be a sequence of independent random variables, uniform on $(0,1)$. Assuming that U and L are independent, let μ denote the law of the pair (L,U). We complete the sequences $(d_{k,n})_{1\leq k\leq H_n}$ and $(\rho_{k,n})_{1\leq k\leq H_n}$ by 0's in order to form the infinite sequences d_n and $\rho^{(n)}$. There are three steps:

(1) we use Theorem 5.2 to derive the convergence of $(L_n, d_n)_{n\geq 1}$ to μ ;
(2) we prove that $\rho^{(n)}$ is the image of (L_n, d_n) by a functional $\mathcal{L}$ whose domain of continuity $\mathcal{C}_{\mathcal{L}}$ satisfies $\mu(\mathcal{C}_{\mathcal{L}}) = 1$;
(3) we check that the distribution of $\mathcal{L}(L,U)$ is conform to the description given in Theorem 1.2.

For step 1, thanks to [Kal97, Theorem 3.29], we know that weak convergence holds for the infinite sequences, if weak convergence of the distribution (under $\mathbb{P}_n$) of the finite sequence $(L_n, (d_{i,n})_{1\leq i\leq k})$ holds for arbitrary k. Again, as in [MZA07, Theorem 6.5], asymptotic independence between $\mathfrak{C}'_n$ and $(d_{i,n})_{1\leq i\leq k}$ holds under $\widehat{\mathbb{G}}_n$: for a $\mathfrak{C}'_n$-measurable $\mathbb{R}$-valued statistic W_n with probability distribution χ_n,

$$\mathcal{W}_2\left((W_n, (d_{i,n})_{1\leq i\leq k}), \chi_n \otimes \mathbb{U}_k\right) = \mathcal{O}\left(\sqrt{\frac{k\log n}{n}}\right). \tag{14}$$

But it holds true, from the definition of L_n, that L_n, or $W_n = e^{-L_n}$, are $\mathfrak{C}'_n$-measurable, i.e. invariant on each $C'(w)$.

Let $\tilde{\chi}_n$ (resp. χ) denote the distribution of W_n under $\mathbb{P}_n$ (resp. the distribution of e^{-L}). Due to Proposition 3.3 and to relation (11),

$$\mathcal{W}_2\left(\tilde{\chi}_n, \chi_n\right) (= \mathcal{W}_2\left(\tilde{\chi}_n \otimes \mathbb{U}_k, \chi_n \otimes \mathbb{U}_k\right)) = \mathcal{O}\left(n^{-1/2+\varepsilon} log n\right).$$

Now, under $\mathbb{P}_n$, L_n has the same law as $L \wedge n$. This is perhaps clearer when one considers the word $\overline{\omega}$ obtained by reading the word ω from right to left : clearly, under $\mathbb{P}_n$, $\overline{L}_n$ defined by

$$\overline{L}_n(\omega) = L_n(\overline{\omega})$$

has the same law as L_n, for ω and $\overline{\omega}$ have the same weight. But $\overline{L}_n$ has the same law as $L \wedge n$. Thus a.s. convergence of $L \wedge n$ to L entails that

$$\mathcal{W}_2\left(\tilde{\chi}_n, \chi\right) (= \mathcal{W}_2\left(\tilde{\chi}_n \otimes \mathbb{U}_k, \chi \otimes \mathbb{U}_k\right)) = \mathcal{O}\left(e^{-n}\right).$$

Weak convergence of $(L_n, (d_{i,n})_{1\leq i\leq k})$ follows at once.

For the point 2, let T be the functional that shifts a sequence u as follows :

$$T(u) = T(u_1, u_2, \dots) = (1, u_1, u_2, \dots).$$

Let S be the functional that keeps track of the sequence of low records (in the broad sense) of a sequence u of real numbers. The functional S is well defined and is continuous on a set of measure 1 of $[0,1]^{\mathbb{N}}$, for instance on the set $\mathcal{R}$ of sequences u without repetition such that $\liminf u = 0$. Then the functional $\mathcal{L}$ defined on $\mathbb{N} \times \mathcal{R}$ by

$$\mathcal{L}(k,u) = T^k \circ S(u)$$

is continuous as well, and $\mathcal{L}(L_n, d_n)$ converges in distribution to $\mathcal{L}(L,U)$.

Set

$$s_{i,n} = 1 - (\rho_{1,n} + \rho_{2,n} + \cdots + \rho_{i,n}).$$

If $L_n(w) = k \geq 1$, the first k factors of the Lyndon decomposition are k words reduced to one letter "a_1". Thus, for $1 \leq i \leq k$,

$$s_{i,n} = 1 - \frac{i}{n}.$$

The next Λ_n terms, $s_{k+1,n}, s_{k+2,n}, \ldots, s_{k+\Lambda_n,n}$, are the low records of the sequence d_n. The difference between the two sequences s_n and $\mathcal{L}(L_n, d_n)$ is thus

$$\mathcal{L}(L_n, d_n) - s_n = \left(\tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{k}{n}, 0, 0, \ldots, 0, s_{k+\Lambda_n+1,n}, s_{k+\Lambda_n+2,n}, \ldots\right).$$

Endowing $[0,1]^{\mathbb{N}}$ with the distance

$$d(u,v) = \sum_{k \geq 1} 2^{-k} |u_k - v_k|,$$

we obtain

$$d(s_n, \mathcal{L}(L_n, d_n)) \leq \frac{L_n^2}{2n} + 2^{-L_n - \Lambda_n}.$$

This inequality and Lemma 5.3 entail that the d-Wasserstein distance between s_n and $\mathcal{L}(L_n, d_n)$ goes to 0. Since $\mathcal{L}(L_n, d_n)$ converges in distribution to $\mathcal{L}(L,U)$, s_n converges in distribution to $\mathcal{L}(L,U)$ too.

For point 3, we note that $\mathcal{L}(L,U)$ and s have the same distribution. Furthermore the transformation sending s_n to $\rho^{(n)}$, and s to ρ, is bicontinuous. Thus the convergence in distribution of $\rho^{(n)}$ to ρ follows.

References

[ABT93] Richard Arratia, A. D. Barbour, and Simon Tavaré, *On random polynomials over finite fields*, Math. Proc. Cambridge Philos. Soc. **114** (1993), no. 2, 347–368. MR MR1230136 (95a:60011)

[ABT99] ———, *On Poisson-Dirichlet limits for random decomposable combinatorial structures*, Combin. Probab. Comput. **8** (1999), no. 3, 193–208. MR MR1702562 (2001b:60029)

[ABT03] ———, *Logarithmic Combinatorial Structures: a probability approch*, European Mathematical Society Zurich, 2003.

[BCN05] Frédérique Bassino, Julien Clément, and Cyril Nicaud, *The standard factorization of Lyndon words: an average point of view*, Discrete Math. **290** (2005), no. 1, 1–25. MR MR2116634 (2005j:68084)

[BD92] Dave Bayer and Persi Diaconis, *Trailing the dovetail shuffle to its lair*, Ann. Appl. Probab. **2** (1992), no. 2, 294–313. MR MR1161056 (93d:60014)

[Bol01] B. Bollobás, *Random graphs*, vol. 73, Cambridge University Press, 2001.

[CG96] J.H. Conway and R.K. Guy, *The book of numbers*, Springer-Verlag, 1996.

[DMP95] P. Diaconis, M.J. McGrath, and J. Pitman, *Riffle shuffles, cycles, and descents*, Combinatorica **15, no. 1** (1995), 11–29.

[GR93] Ira M. Gessel and Christophe Reutenauer, *Counting permutations with given cycle structure and descent set*, J. Combin. Theory Ser. A **64** (1993), no. 2, 189–215. MR MR1245159 (95g:05006)

[Han93] Jennie C. Hansen, *Factorization in* $\mathbf{F}_q[x]$ *and Brownian motion*, Combin. Probab. Comput. **2** (1993), no. 3, 285–299. MR MR1264035 (95f:11056)

[Han94] ———, *Order statistics for decomposable combinatorial structures*, Random Structures Algorithms **5** (1994), no. 4, 517–533. MR MR1293077 (96f:60010)

[Kal97] O. Kallenberg, *Foundations of Modern Probability*, Springer series in Statistics Probability and its applications, 1997.

[Kin75] J. F. C. Kingman, *Random discrete distributions*, Journal of the Royal Statistical Society. Series B (Methodological) **37** (1975), no. 1, 1–22.

[Lot83] M. Lothaire, *Combinatorics on words*, vol. 17, Encyclopedia of mathematics and its applications, Addison-Wesley, 1983.

[Lot02] ———, *Algebraic Combinatorics on Words*, vol. 90 of Encyclopedia of mathematics and its applications, Cambridge University Press, 2002.

[Lot05] ———, *Applied Combinatorics on Words*, Cambridge University Press, 2005.

[Lyn54] R. Lyndon, *On Burnside problem I*, Trans. American Math. Soc. **77** (1954), 202–215.

[McC65] J.W McCloskey, *A model for the distribution of individuals by species in an environment*, Unpublished Ph.D. thesis, Michigan State University (1965).

[MZA07] R. Marchand and E. Zohoorian Azad, *Limit law of the length of the standard right factor of a Lyndon word*, Combin. Probab. Comput. **16** (2007), no. 3, 417–434. MR MR2312436 (2008e:68120)

[Oka58] M. Okamoto, *Some inequalities related to the partial sum of binomial probabilities*, Ann. Inst. Statist. Math. **10** (1958), 29–35.

[PPY92] Mihael Perman, Jim Pitman, and Marc Yor, *Size-biased sampling of Poisson point processes and excursions*, Probab. Theory Related Fields **92** (1992), no. 1, 21–39. MR MR1156448 (93d:60088)

[Rac91] S.T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, U.K., 1991.

[Reu93] C. Reutenauer, *Free lie algebras*, Oxford Science Publications, 1993.

[SW09] G.R. Shorack and J.A. Wellner, *Empirical processes with applications to statistics*, Society for Industrial Mathematics, 2009.

Institut Elie Cartan Nancy (mathématiques), Université Henri Poincaré Nancy 1, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex France

E-mail address: Philippe.chassaing@iecn.u-nancy.fr

Mathematics faculty, Damghan University of Basic Sciences, Damghan, Iran

E-mail address: Elahe.Zohoorian@iecn.u-nancy.fr